A region-level graph labeling approach to motion-based segmentation

Marc Gelgon, Patrick Bouthemy

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

A region-level graph labeling approach to motion-based segmentation

Marc GELGON, Patrick BOUTHEMY

N° 3054

decembre 1996

THÈME 3

ISSN 0249-6399

INRIA
RENNES

A region-level graph labeling approach to motion-based segmentation

Marc GELGON, Patrick BOUTHEMY

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet Temis

Rapport de recherche n°3054 — decembre 1996 — 19 pages

Abstract:

This paper deals with the problem of motion-based segmentation of image sequences. Such partitions are multiple-purpose in dynamic scene analysis. We first extract a texture-based partition using an unsupervised MRF approach. The regions obtained are then grouped according to a motion-based criterion. This grouping process relies on two motion estimation techniques and exploits contextual information between regions. In contrast with clustering techniques, region grouping is formalized as a motion-based graph labeling process, within a Markovian framework. Results on real-world image sequences are shown and validate the proposed method.

Key-words: Image sequence, Markovian models, segmentation, motion analysis, region grouping, graphs, statistical labeling.

(Résumé : tsvp)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 02 99 84 71 00 – Télécopie : (33) 02 99 84 71 71

Segmentation en régions au sens du mouvement apparent 2D par étiquetage statistique de graphes.

Résumé : Dans ce rapport, nous nous intéressons à la segmentation en régions selon un critère de mouvement 2D dans des séquences d'images . Ce type de partition peut être exploité à de nombreuses fins en analyse de scène dynamique. Dans la méthode proposée, une partition spatiale (texturelle) est d'abord extraite grâce à une approche Markovienne non supervisée. Les régions obtenues sont ensuite regroupées selon un critère de mouvement. La méthode de regroupement repose sur deux techniques d'estimation du mouvement et exploite une information contextuelle au niveau des régions. Le regroupement est formulé comme un problème d'étiquetage des noeuds d'un graphe d'adjacence des régions spatiales et est effectué dans un cadre Markovien. Des résultats sur des images réelles sont fournis et forment une première validation de la méthode.

Mots-clé : Séquence d'images, modèles Markoviens, segmentation, mouvement, regroupement de régions, graphes, étiquetage statistique.

1 Introduction

Motion-based segmentation of images is among the important tasks in the field of computer vision, since the image components thereby extracted generally correspond to meaningful entities. Provided they can be obtained for a whole image sequence, such partitions can serve as data input for region-based coding schemes[15], tracking procedures[10] or interpretation stages of the dynamic content of the observed scene [4].

A first category of approaches is based on building up the motion-based regions directly as groupings of pixels. Among these methods, the top-down techniques consist in computing successive estimations of dominant motions using for instance a least-squares technique [8]. Significant region areas that do not conform to the dominant motion model make up new regions, and the process is iterated. In [16], the parameters of the models describing the various motions in the frames are first estimated by searching for minima in parameter space with an intensity-matching criterion. The segmentation is then achieved by associating pixels to the motion model leading to the smallest displaced frame difference. Other methods carry out a joint estimation of the motion models and of their spatial support. In [7], an unsupervised clustering technique mixes information on the position, color and motion-based residual at every pixel in a competitive learning scheme, exploiting assumed Gaussian distributions of these features. A Markovian framework exploiting a robust estimator was chosen in [11]. In [1], motion models and spatial supports are simultaneously estimated as the two steps of an EM algorithm, whereas the number of motion models is jointly determined using a MDL-based criterion. In [13], the problem of simultaneous estimation of the segmentation map and of the dense motion field is addressed in a Markovian framework using a generalized Gaussian model of the prediction error distribution and constraints on the motion field and the segmentation map.

Another class of methods introduces a layer of intermediate regions, built from pixels according to a determined procedure, which form the basic elements of the definitive regions. Our work enters this category. Three main characteristics of these schemes can be distinguished : the homogeneity criterion within intermediate regions, the region merging (possibly splitting) criterion, and the procedure carried out to perform this merging. Intermediate regions can display motion [17],[14] and [18], or intensity homogeneity properties [6],[15] and [2]. Concerning the region grouping technique to be considered, clustering algorithms and region-level Markovian models are among the main frameworks proposed. The first category includes k-means [14], k-medoid [6] and more elaborate techniques [15] involving similarity

measures between motion model parameters [14] or displaced frame difference residuals [15]. A MDL criterion is used in [18] to decide upon the merging of two regions. An advantage of the MDL criterion is that having to set a threshold upon a motion discrepancy measure can be avoided. However, the coding optimality principle is not guaranteed to match dynamic analysis purposes. In contrast to the clustering approaches, an algorithm employing a region-level Markovian model is proposed in [17].

In our case, the motion-based segmentation is associated to a region-level motion-based valued graph. The combination of a spatial (texture-based) partition with this graph carries a description of the partition formed of both pixel-level and region-level information. Because an accurate determination of object contours is generally easier using a intensity criterion rather than a motion-based one, one can expect from a scheme combining both criteria some noticeable improvement over motion-only-based methods. Moreover, this representation is embedded in a partition prediction-update scheme along the sequence.

The remainder of this paper is organized as follows. An overview of the segmentation method is presented in Section 2. Section 3 outlines the texture-based algorithm. Section 4 details how a region-level graph can be built and labeled. Section 5 contains a set of results obtained on real-world sequences. Conclusions are drawn in Section 6.

2 Overview of the motion-based segmentation method

Fig.1 presents an outline of the segmentation technique we propose. Given two successive frames at time t and $t+1$, we aim at extracting a spatio-temporal partition of the frame at t. Three kinds of partitions are successively involved :

- a texture-based partition of the image (fig. 1.a)
- a motion-based labeling of a graph (fig. 1.c)
- a spatio-temporal (motion-based) partition of the image (fig. 1.d)

An intensity-based segmentation of the image is first required. It is supplied by a MRF-based texture segmentation method, described in the next section. A spatial graph is derived from the topology of this texture-based segmentation (fig 1.b.). The regions, corresponding to nodes on the graph, are then to be grouped according to a motion homogeneity criterion. To this end, each node is assigned a label identifying the grouping it is associated to. We search for a label configuration such that

regions undergoing similar (resp. different) motions are attributed the same (resp. different) labels. To this aim, we introduce a MRF model, which sites are the nodes of the graph. Motion is estimated within each region of the spatial partition, using a robust parametric estimator relying on a differential method. Should the motion estimates not be reliable enough, according to the uncertainty measurement provided by this estimator, motion is estimated alternatively by a matching technique based on a similarity transform estimation with a stochastic algorithm. The label configuration is optimized through the definition of an appropriate energy function and its minimization. Through the definition of the energy function, region-level contextual information can be introduced to regularize the label estimation problem, which clustering techniques cannot easily incorporate in a well formalized manner. We wish to point out that, in contrast to region clustering approaches in which region merging is mostly irreversible, the principle of graph labeling in a Markovian framework enables to challenge groupings through the search for an optimal configuration. Finally, a motion-based partition of the image can be inferred from the label-based partition of the graph.

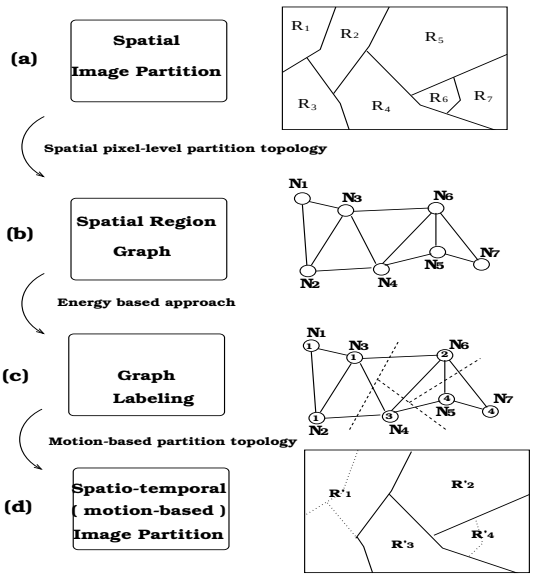


Figure 1: *Overview of the algorithm : how the spatio-temporal partition is derived from a texture-based partition.*

As in other works (e.g.[2]), we shall assume a unique motion per intermediate region, i.e. motion boundaries coincide with spatial boundaries. This assumes a sufficiently fine spatial segmentation, which in turn implies that the problem of motion estimation is particularly cared for, to cope with regions with possibly poor spatio-temporal gradient information.

Both energy minimization problems related to the pixel-level and region-level Markovian models can benefit from partition predictions available at these levels. In both cases, the availability of a label map close to the optimal one enables the use of a rapidly converging deterministic relaxation algorithm, provided no special event (such as strong occlusion or crossing) occurs. Moreover, the labeling associated to both the spatial and spatio-temporal partition remain naturally coherent from t to $t+1$, through the initialization of the relaxation scheme by the predicted motion-oriented label map. This is of key importance for defining an efficient tracking procedure.

3 Unsupervised MRF-based texture segmentation

The spatial segmentation is performed by applying the unsupervised technique described in [9]. The principles of the method are outlined in this section.

It operates within a Bayesian estimation framework. Let $e = \{e_s, s \in S\}$ (resp. $o = \{o_s, s \in S\}$) be a realization of a random label field E (resp. of a random observation field O), over a set of sites S corresponding to the image pixels. Given a neighbourhood system, (E, O) is modeled as a Markov Random Field. Using the Gibbs distributions / MRF equivalence, the optimal label field $\widehat{e}$ is derived according to the *Maximum a posteriori (MAP)* criterion, and is expressed as follows: $\widehat{e} = \arg\min_{e \in \Omega} U(e, o)$, where Ω is the set of all possible realizations of E and $U(e, o)$ is the so-called energy function encompassing the interactions between labels and observations and a priori information on the label field.

For most real-world images, the observation vector $\{o_s = [o_s^1, \ldots, o_s^m], s \in S\}$ can be composed of only two simple features : grey-level and local contrast. For images including regions with more elaborate textures, other more discriminating features can be added.

The energy function is defined as follows :

$$U(e, o) = U_1(e, o) + U_2(e) \tag{1}$$

The data-driven term U_1 decomposes as a sum of local potentials : At site s, the distribution of each texture feature is computed within a local window centered

at s. The likelihood of this site being assigned the label e_s is determined by the Kolmogorov-Smirnov distances $d(.,.)$ between the local distributions and the feature distributions computed on region R_{e_s}.
The local potentials are related linearly to this distance according to predetermined constants $c^{(i)}$.

$$U_1(e,o) = \sum_{s \in S} \sum_{i=1}^{m} c^{(i)} d(o^{(i)}(R_{e_s}), o^{(i)}(B_s)) \tag{2}$$

$$U_2(e) = \sum_{(s,t) \in \mathcal{C}} \mu(1 - 2\delta(e_s - e_t)) \tag{3}$$

μ is a predetermined positive constant, and $\delta(e_s - e_t)$=1, if $e_s = e_t$, 0 otherwise. This term favours spatial homogeneity of regions.

Energy minimization is performed using a modified HCF algorithm[5]. At a given site, in addition to the current label and to the labels assigned in the neighbourhood, an outlier label ρ is among the candidate labels [4],[9]. A parameter ϕ sets the local energy variation associated to this label and hence indirectly the number of regions created. Once the relaxation process is completed, new labels are attributed to the connected subsets of sites with the ρ-label, which size exceeds a pre-set threshold.

An on-line determination of the number of regions can thereby be achieved. If a predicted segmentation map is available, the region borders are adjusted and regions are, if necessary, automatically created (or removed). Should no information about the label map be available, the segmentation map is initially considered as one region and updated using a maximum likelihood estimator. This segmentation algorithm has the major advantage of being unsupervised, both as far as the number of regions and their texture characteristics are concerned.

4 Region-level graph building and labeling

The structure exploiting the spatial partition to build a motion-based segmentation is the topic of this section. Let $\mathcal{G}$ be the adjacency graph derived from the topology of the spatial partition $\mathcal{P} = \{R_k, k = 1 \ldots p\}$, containing p regions. We denote N_k its nodes, which correspond to the regions R_k of the spatial partition. Let arcs A_j joint in $\mathcal{G}$ the nodes associated to neighbouring regions in the spatial partition.

$$\mathcal{G} = \{\{N_1, \ldots, N_p\}, \{A_1, \ldots, A_q\}\} \tag{4}$$

We aim at assigning a motion label to every node in the graph, with a view to partitioning this graph into node subsets corresponding to groupings of regions

of coherent motion. The labeling of the graph is formalized within a Markovian framework. To this purpose, we identify the nodes of the graph to the sites of a region-level MRF thus defined on an irregular grid. A pair of neighbouring sites on the grid is called a clique. The cliques are deduced in a straightforward manner from the arcs of the graph. Let $\nu = \{\nu_1, \ldots, \nu_p\}$ be the set of sites and $\Gamma = \{\gamma_1, \ldots, \gamma_q\}$ be the set of binary cliques. Let $e'(\nu_k)$ be the label attached to node ν_k, which identifies which grouping the associated spatial region belongs to. We now focus on the definition of a suitable energy function for our motion-based region grouping objective.

4.1 Energy function definition

As in the case of the pixel-level energy function for the texture-based segmentation stage, the region-level energy function U' can be split up into a observation/label interaction term and a regularization term. However, the data-driven term is here defined over a binary clique. The energy function is expressed as :

$$U'(e', o') = \sum_{\gamma_j \in \Gamma} V_1'(e'(\gamma_j), o'(\gamma_j)) + \sum_{\gamma_j \in \Gamma} V_2'(e'(\gamma_j)) \tag{5}$$

where $e'(\gamma_j)$ stands for the pair of labels attached to the clique γ_j ($\gamma_j = \{\nu_k, \nu_{k'}\}$) and o' for the region-level observations, which we will examine below. Potential V_1' is equivalent to a discrepancy measure between the two motion model fields attached to the sites γ_k and $\gamma_{k'}$ composing clique γ_j. The motion model estimation techniques and the chosen discrepancy measure are now presented.

4.2 Motion estimation

The inter-frame transformation between frame I_t at time t and frame I_{t+1} at time $t+1$ is modeled by a set of 2D affine motion models, one per region. Differential (gradient-based) motion estimation techniques can determine effectively such a model when applied on sufficiently textured regions. On poorly-textured regions, however, resorting to an intensity-matching approach generally leads to more reliable estimates. We estimate the motion model for each region R_k on the partition as follows. The gradient-based motion estimator is first applied to each region.

Differential motion estimator

The displacement vector at site $s = (x, y)$ in region R_k which gravity center $g_k = (x_g^k, y_g^k)$, is expressed as:

$$\overrightarrow{d_{(\Theta_k)_t^{t+1}}}(s) = \begin{pmatrix} a_0^k + a_2^k(x - x_g^k) + a_3^k(y - y_g^k) \\ a_1^k + a_4^k(x - x_g^k) + a_5^k(y - y_g^k) \end{pmatrix} \tag{6}$$

in which the motion parameter vector $(\Theta_k)_t^{t+1} = [a_0^k \dots a_5^k]$ is estimated on each region
$R_k, k = 1 \dots p$, using the robust multi-resolution estimator called *RMRmod* described in [12]. It aims at minimizing the following criterion :

$$\sum_{s \in R_k(t)} \rho\left(\mathrm{DFD}(s, \Theta_k)\right) \tag{7}$$

where $DFD(s, \Theta_k) = I_{t+1}(s + \overrightarrow{d_{(\Theta_k)_t^{t+1}}}(s)) - I_t(s)$ and $\rho()$ is Tukey's function. Because $DFD(s, \Theta_k)$ is not linear with regard to Θ_k, we operate with an incremental strategy. The initial value of the estimate $\widehat{\Theta_k}$ can be set to null or derived, for instance, from motion information obtained from previous frames. A succession of estimate refinements $\Delta\Theta_k$ are computed (and cumulated) using successive first order approximations r_s of the residual $DFD(s, \Theta_k)$, which are linear with regard to the increment $\Delta\Theta_k$ to be estimated. Denoting $\widehat{\Theta_k}$ the sum of increments computed so far, the residual r_s is expressed as :

$$\begin{aligned} r_s &= I(s + \overrightarrow{d_{(\widehat{\Theta_k})}}(s), t+1) - I(s,t) \\ &\quad + \nabla I(s + \overrightarrow{d_{(\widehat{\Theta_k})}}(s), t+1).\overrightarrow{d_{(\Delta\Theta_k)}} \end{aligned} \tag{9}$$

At each iteration, we search for :

$$\widehat{\Delta\Theta_k} = \arg\min_{\Delta\Theta_k} \sum_{s \in R_k(t)} \rho\left(r_s(\Delta\Theta_k)\right)) \tag{10}$$

Increments are cumulated to make up the estimate $\widehat{\Theta_k}$ until a predefined convergence criterion is met. This scheme exploits an Iterative Weighted Least Squares

procedure, and in fact only involves the computation of the spatio-temporal derivatives of the intensity function. Owing to the robustness of the estimator, the motion measurement is rather insensitive to minor errors in region border determination and to motions due to small mobile objects, if any within the region.

An estimation of the covariance matrix associated to the motion parameter vector is also provided.The diagonal elements of the covariance matrix associated to the motion vector $\overrightarrow{d_{\Theta_k}} = (d_x(s), d_y(s))$ are expressed as :

$$\sigma^2_{d_x} = \sigma^2_{a^0_k} + \sigma^2_{a^1_k}(x - x^k_g)^2 + \sigma^2_{a^2_k}(y - y^k_g)^2 \tag{11}$$

$$\sigma^2_{d_y} = \sigma^2_{a^3_k} + \sigma^2_{a^4_k}(x - x^k_g)^2 + \sigma^2_{a^5_k}(y - y^k_g)^2 \tag{12}$$

We set the maximum tolerance on the mean estimation error on the displacement vector, over the region, to δ pixels. In practise δ is set to 3. Assuming a confidence interval of $3\sigma_{d_x}$ and $3\sigma_{d_y}$ around d_x and d_y, the estimated parameter vector is considered to be reliable if the following test is passed :

$$\left[\frac{1}{card(R_k)} \sum_{s \in R_k(t)} (\sigma^2_{d_x} + \sigma^2_{d_y})\right] < 2\left(\frac{\delta}{3}\right)^2 \tag{13}$$

In the case the threshold is exceeded, another estimation of a more simple motion model is carried out, exploiting directly the intensity values rather than the intensity spatio-temporal gradients.

Intensity matching motion estimator

The model to be estimated is a 4-parameter simplified affine model, allowing description of translation, rotation and divergence.

$$\overrightarrow{d_{(\Theta_k)^{t+1}_t}}(s) = \begin{pmatrix} a^k_0 + a^k_2(x - x^k_g) - a^k_3(y - y^k_g) \\ a^k_1 + a^k_3(x - x^k_g) + a^k_2(y - y^k_g) \end{pmatrix} \tag{14}$$

Let us denote α the rotation angle and κ the divergence ratio of the geometrical transform associated to the model. The criterion to be minimized is a simple Displaced Frame Difference. An optimal vector is sought for, using a stochastic Gibbs sampler, in the parameter space, which is irregularly discretized in terms of a_0, a_1, α and κ. This simulated annealing process is performed in a search space of moderate size so that it does not overload the computation. Moreover, such regions generally have a small size and represent a minority of spatial regions.

4.3 Construction of a motion-based distance between regions

In order to characterize the difference between the estimated motions within two neighbouring regions R_i and R_j, we consider the two motion fields issued from the motion models estimated within each region. We extend these fields over the support corresponding to the union of the two regions. The discrepancy between these two fields, denoted $D(\gamma_k)$, is expressed as the average, over the union of the two regions, of a weighted distance ϵ between the velocity vectors that form these fields :

$$D(\gamma_k) = \frac{1}{card(R_i \cup R_j)} \sum_{s \in (R_i \cup R_j)} \epsilon(\overrightarrow{d_{\Theta_i}}(s), \overrightarrow{d_{\Theta_j}}(s)) \tag{15}$$

We wish to define a distance between motion vectors taking into account the covariance information.

Much investigation has been devoted to the case of distances between two Gaussian vectors [3]. Measures such as Bhattacharya or Malahanobis distance, or Kullback divergence, are well suited to measure separability between two distributions. From the dynamic range of variances measured on many regions and frames, we noticed that these types of distances ($\frac{1}{\sigma^2}$ laws) had too strong an influence on the distance. A log-type weighting function with moderate and well-controlled weight was hence prefered. Such a weighting function was defined as follows :

$$\begin{aligned} \epsilon(\overrightarrow{d_{\Theta_i}}, \overrightarrow{d_{\Theta_j}}) = & |d_{x,j} - d_{x,i}|.f(\sigma^2_{d_{x,j}} + \sigma^2_{d_{x,i}}) \\ & + |d_{y,j} - d_{y,i}|.f(\sigma^2_{d_{y,j}} + \sigma^2_{d_{y,i}}) \end{aligned} \tag{16}$$

If the estimates from gradient-based method are retained, $f(x) = -0.15.\log_{10}(x) - 0.2$. These constants were set according to the motion vector variance dynamic range experimentally observed, so that for $(\sigma^2_{d_{y,j}} + \sigma^2_{d_{y,i}}) = 10^{-8}$, $f(x) = 1$ and for $(\sigma^2_{d_{y,j}} + \sigma^2_{d_{y,i}}) = 10^{-2}$, $f(x) = 0.1$. If the intensity-matching method is used, we take $f(x) = 0.3$.

V_1' (fig. 2) aims at favouring identical neighbouring labels when the attached motions are similar, and different labels when motions are strongly different. In contrast with the binary penalty value resulting from a test on the hypothesis that two estimated motions really correspond to two really identical underlying motions defined in [17], a progressive transition is introduced here. The potential is defined in (17) below. Only one parameter (τ) sets both the threshold and the slope of the function.

$$V_1'(e'(\gamma_k), o'(\gamma_k)) = \begin{cases} \frac{1}{1+e^{\frac{2}{\tau}(D(\gamma_k)-\tau)}} & \text{if } e_i' = e_j' \\ 1 - \frac{1}{1+e^{\frac{2}{\tau}(D(\gamma_k)-\tau)}} & \text{if } e_i' \neq e_j' \end{cases} \quad (17)$$

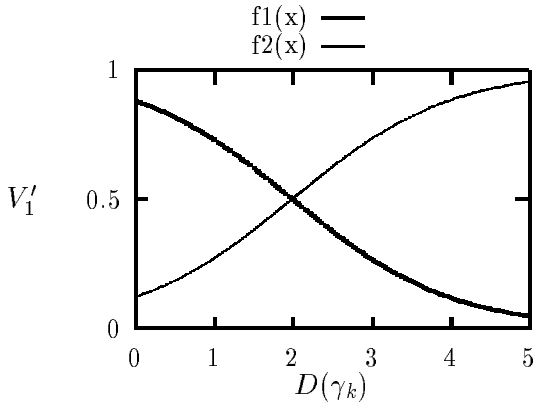


Figure 2: *V_1' potential as a function of the difference $D(\gamma_k)$ between two motion fields, for identical labels (curve f2) and different labels (curve f1). Parameter value : $\tau = 2$.*

V_2' corresponds to the regularization term. It aims at favouring identical labeling of two neighbouring regions, taking into account the "degree" of adjacency between two regions. Because it is defined, as V_1', on the binary cliques of the field, it can be viewed as a term that backs the data-driven potential, depending on geometrical characteristics of the region pair. Two geometrical features are computed per region pair (R_i, R_j): the length of the common border (denoted as $\xi_{i,j}$) and the distance between the region gravity centers g_i and g_j. They are combined into a geometrical "compacity factor" $\eta_{i,j}$ of the pair which takes part in the definition of the potential:

$$\eta_{i,j} = \frac{\xi_{i,j}}{\xi_{i,j}+||g_i - g_j||_2} \quad (18)$$

$$V_2'(e'(\gamma_k)) = \begin{cases} -\beta.\eta_{i,j}, \quad \beta > 0 & \text{if } e_i' = e_j' \\ 0 & \text{if } e_i' \neq e_j' \end{cases} \quad (19)$$

The relative small number of regions allows us to utilize an energy minimization technique based on the HCF method [5]. For the first frame of the sequence, all

regions are initially given different motion labels. Sites are visited according to their rank in an unstability stack [5]. Candidate labels at a given site include the current motion label at this site and the motion labels currently assigned to the neighbouring sites, as well as an extra label. For each candidate label, the local energy involved is computed according to definitions (17) and (19) above. The motion label giving rise to the greatest local energy descent is assigned to the node. If it is the extra label, this corresponds to creating a new motion-based region on the pixel-level partition. This extra label can thereby lead the optimization process out of undesirable local minima and enable a correct on-line determination of the number of relevant motion entities.

In region-merging approaches, the determination of region groupings generally corresponds to a progressive graph reduction (topology simplification). Our approach dissociates the graph topology and the constitution of region groupings through the node labeling. A major advantage of formulating a region-level grouping in the proposed framework is that if two or more regions become grouped through some energy minimization steps, they can later be labeled differently by some further optimization steps, if such a configuration leads to a lower global energy.

5 Results

This section presents three sets of results corresponding to motion-based segmentation of three real-world sequences.

The parameters of the algorithm were set as follows. For the pixel-level segmentation, ϕ=0.5, μ=0.05 and $c^{(i)}$=2 for all sequences. For the graph labeling, τ was set to 3 for $Interview$ and $Renata$ and to 10 for Van. β was set to 0.2 for all sequences.

In the $Interview$ sequence (fig. 3), the woman is getting up while twisting slightly clockwise around the vertical axis of her body. This leads to a complex articulated 2D motion. Meanwhile, the background is moving in the same direction as the woman due to the movement of the camera, but more slowly. The spatial-based and motion-based segmentation contours are superimposed on the original images and are respectively shown in fig. 3(a) and fig. 3(b) for frame 13, and in fig. 3(c) and fig. 3(d) for frame 19. It can be seen on the motion-based partitions that errors are mainly caused by some regions of the spatial partition breaking the assumption of unique motion per spatial region. Taking into account that the knees are rising more slowly than the rest of the body because of the articulated nature of the motion, the motion contours are overall well delimited.

In the *Van* sequence (fig. 4), the van is going from left to right and slightly inwards the image. A white car is driving from left to right, passing behind the van as seen from the camera. The background is static. The spatial-based and motion-based segmentation contours are respectively shown on fig. 4.a and fig. 4.b, for frame 28, and on fig. 4.c and fig. 4.d, for frame 37. Both the car and the van are correctly separated from the background. The van is well delimited on most of its contour. The area corresponding to the sky undergoes strong and spatially irregular noise. This leads to unstable motion estimates, hence the segmentation is perturbed in this area.

The woman in the *Renata* sequence (fig. 5) is going right, while the camera is going in the same direction, but more slowly. Her right is arm slightly swinging. The spatial-based and motion-based segmentation contours are respectively shown on fig. 5.a and fig. 5.b, fig. 5.c and fig. 5.d, for frames 5 and 21. On frame 5, the arm is grouped with the calendar on the background, their apparent motion being similar because the arm is then swinging backwards. Her arm then slowly swings back forward. Thus, the arm and the body moving similarly are grouped on frame 21.

In order to alleviate the problem of non-unique motion per spatial region, and rather than increasing ϕ to obtain more spatial regions, which would create many small trouble-causing regions with uniform intensity, we plan to exploit the motion detection algorithm on which [11] is based to split, when necessary, spatial regions.

6 Conclusion

A method for motion-based segmentation was presented, relying on a hierarchical segmentation scheme of the image. A pixel-level texture-based segmentation of the image was first obtained employing a Markovian approach. A region-level adjacency graph was built on the partition obtained. The derivation of a motion-based partition of the image was achieved through a graph labeling process in a Markovian framework. To this aim, an appropriate energy function was defined.

The graph-labeling principle provides advantages over classical merging methods which, by operating a graph reduction, imply irreversibility of merging. The energy-based approach avoids critical dependency on the order in which regions are merged. Moreover, since the motion-based segmentation retains information about both the spatial partition and the label configuration, both require only updating from one frame to the next, thanks to the Markovian framework. Promising results have been obtained so far, requiring no fine parameter tuning.

The framework presented opens several extension opportunities. Partition tracking can be carried out by estimating motion models on motion-based region groupings and then building a motion-oriented prediction map. This could be coupled to a recursive filtering scheme of the motion model parameters. We are also currently dealing with the enhancement of region primitive description. By means of a geometrical description of regions with an associated long-term tracking process, we aim at coping efficiently with occlusions and crossings. We plan to define a method that handles the tracking the region partition as a whole (regions are jointly tracked) rather than a given particular region.

References

[1] S. Ayer and H.S. Sawhney. Layered representation of motion video using a robust maximum-likelihood estimation of mixture models and MDL encoding. In *Proc of IEEE Int. Conf. on Computer Vision*, pages 777–784, MIT Cambridge, MA, June 1995.

[2] S. Ayer, P. Schroeter, and J. Bigün. Segmentation of moving objets by robust motion parameter estimation over multiple frames. In *Proc. of Third European Conference on Computer Vision*, pages 316–327, Stockholm, Sweden, May 1994.

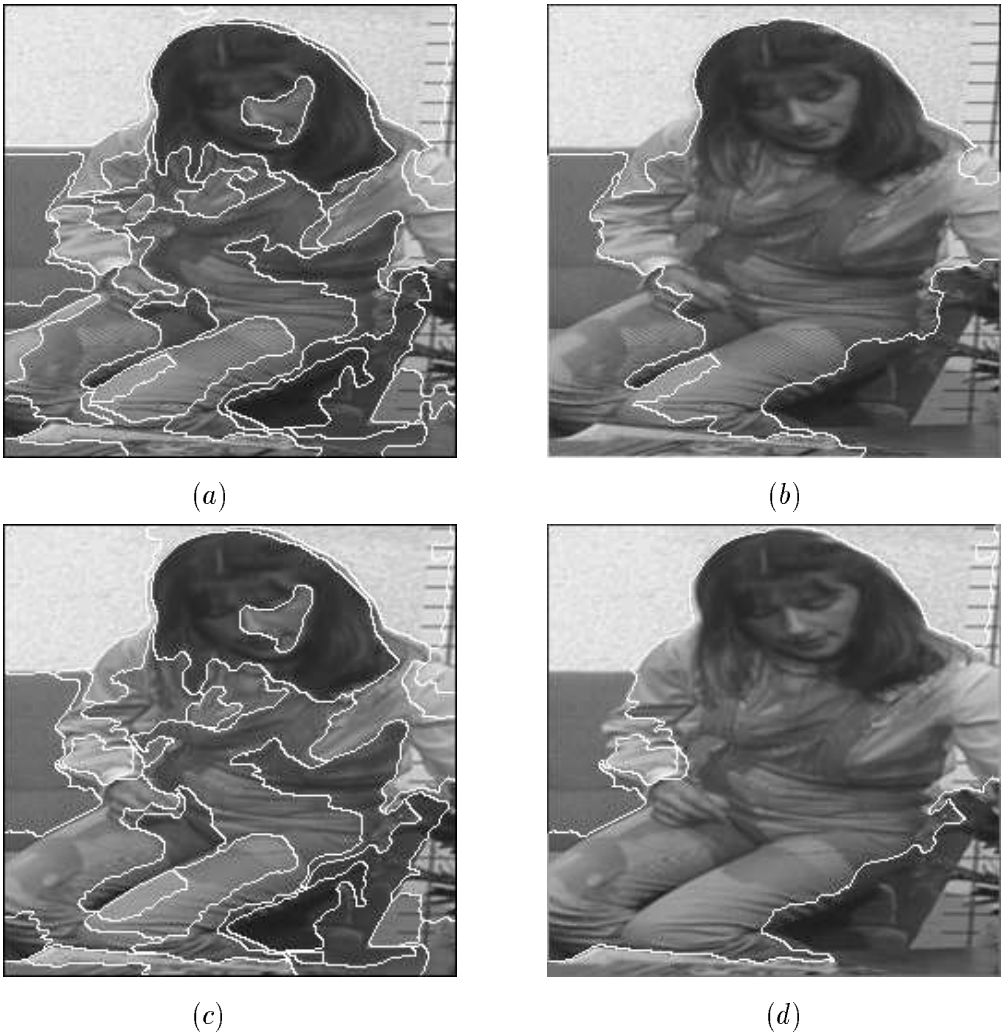
[3] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, December 1989.

[4] P. Bouthemy and E. François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *International Journal of Computer Vision*, 10(2):1578–182, April 1993.

[5] P.B. Chou and C.M. Brown. The theory and practise of Bayesian image modelling. *International Journal of Computer Vision*, 4:185–210, 1990.

[6] F. Dufaux, F. Moscheni, and A. Lippman. Spatio-temporal segmentation based on motion and static segmentation. In *Proc of Second Int. Conf. of Image Processing*, pages 306–309, Washington, October 1995.

[7] M. Etoh and Y Shirai. Segmentation and 2D motion estimation by region fragments. In *Proc of IEEE Int. Conf. on Computer Vision*, pages 192–199, Berlin, May 1993.



(*a*) (*b*)

(*c*) (*d*)

Figure 3: *Interview* sequence : the spatial-based and motion-based segmentation maps for frame 13 (a)(b) and frame 19 (c)(d).

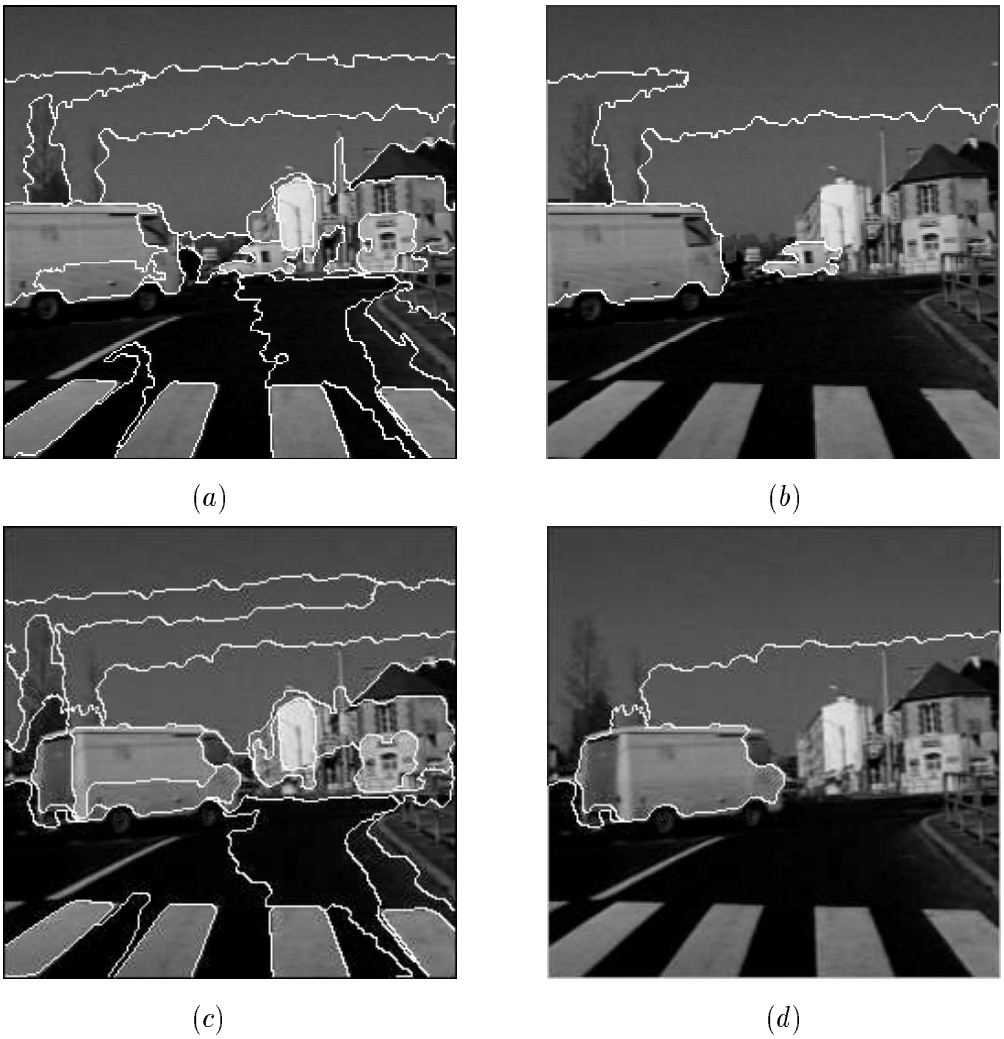


Figure 4: *Van* sequence : the spatial-based and motion-based segmentation maps for frame 28 (a)(b) and frame 37 (c)(d).

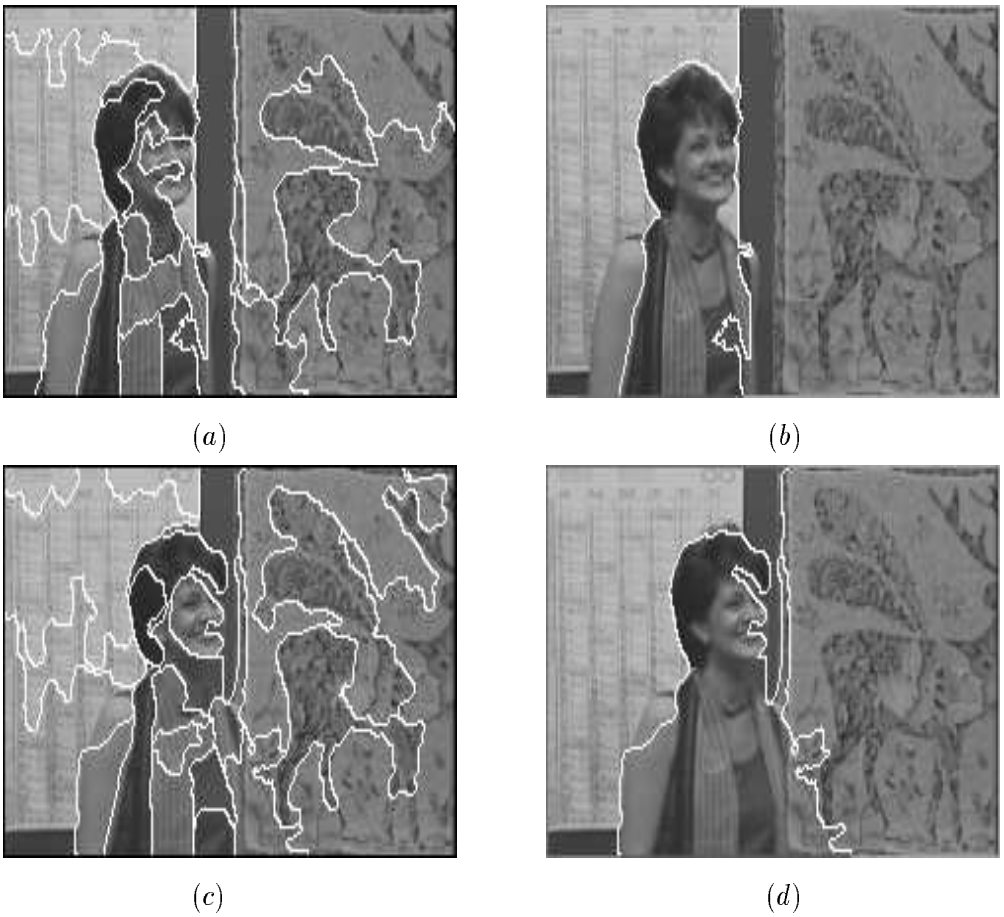


Figure 5: *Renata* sequence : the spatial-based and motion-based segmentation maps for frame 5 (a)(b) and frame 21 (c)(d)

[8] M. Irani, B. Rousso, and S. Peleg. Detecting and tracking multiple moving objects using temporal integration. In *Proc. of Second European Conference on Computer Vision*, pages 282–287, Santa Margherita Ligure, Italy, May 1992.

[9] C. Kervrann and F. Heitz. A Markov random field model-based approach to unsupervised texture segmentation using local and global spatial statistics. *IEEE Trans. on Image Processing*, 4(6):856–862, June 1995.

[10] F. Meyer and P. Bouthemy. Exploiting the temporal coherence of motion for linking partial spatio-temporal trajectories. In *Proc of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, pages 746–747, New-York, 1993.

[11] J.M Odobez and P. Bouthemy. MRF-based motion segmentation exploiting a 2D motion model robust estimation. In *Proc of Second Int. Conf. of Image Processing*, pages 628–631, Washington, October 1995.

[12] J.M Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, 6(4):348–365, December 1995.

[13] C. Stiller. Object-based motion computation. In *Proc. of Int. Conf. on Image Processing*, pages 913–916, Lausanne, Sept 1996.

[14] J.Y.A Wang and E.H Adelson. Representing moving images with layers. *IEEE Trans. on Image Processing*, 3(5):625–638, September 1994.

[15] L. Wu, J. Benois-Pineau, Ph. Delagnes, and D. Barba. Spatio-temporal segmentation of image sequences for object-oriented low bit-rate image coding. *Signal Processing : Image Communication*, 8:513–543, September 1996.

[16] S.F. Wu and J. Kittler. A gradient-based method for general motion estimation and segmentation. *Journal of Visual Communication and Image Representation*, 4(1):25–38, March 1993.

[17] W. Xiong and C. Graffigne. A hierarchical method for detection of moving objects. In *Proc of First Int. Conf. of Image Processing*, pages 795–799, Austin, November 1994.

[18] H Zheng and D. Blostein. Motion-based object segmentation and estimation using the MDL principle. *IEEE Trans. on Image Processing*, 4(9):1223–1235, September 1995.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399